Agnostic feature selection

Guillaume Florent Doquet, Michèle Sebag

Agnostic feature selection

Guillaume Doquet ✉ and Michèle Sebag

TAU
CNRS − INRIA − LRI − Université Paris-Saclay, France
{doquet,sebag}@lri.fr

Abstract. Unsupervised feature selection is mostly assessed along a supervised learning setting, depending on whether the selected features efficiently permit to predict the (unknown) target variable. Another setting is proposed in this paper: the selected features aim to efficiently recover the whole dataset. The proposed algorithm, called AgnoS, combines an AutoEncoder with structural regularizations to sidestep the combinatorial optimization problem at the core of feature selection. The extensive experimental validation of AgnoS on the scikit-feature benchmark suite demonstrates its ability compared to the state of the art, both in terms of supervised learning and data compression.

Keywords: clustering and unsupervised learning, feature selection, interpretable models

1 Introduction

With the advent of big data, high-dimensional datasets are increasingly common, with potentially negative consequences on the deployment of machine learning algorithms in terms of i) computational cost; ii) accuracy (due to overfitting or lack of robustness related to e.g. adversarial examples (Goodfellow et al., 2015)); and iii) poor interpretability of the learned models.

The first two issues can be handled through dimensionality reduction, based on feature selection (Nie et al., 2016; Chen et al., 2017; Li et al., 2018) or feature construction (Tenenbaum et al., 2000; Saul and Roweis, 2003; Wiatowski and Bölcskei, 2018). The interpretability of the learned models, an increasingly required property for ensuring *Fair, Accountable* and *Transparent* AI (Doshi-Velez and Kim, 2017), however is hardly compatible with feature construction, and feature selection (FS) thus becomes a key ingredient of the machine learning pipeline.

This paper focuses on *unsupervised feature selection*. Most FS approaches tackle supervised FS (Chen et al., 2017), aimed to select features supporting a (nearly optimal) classifier. Quite the contrary, unsupervised feature selection is not endowed with a natural learning criterion. Basically, unsupervised FS approaches tend to define pseudo-labels, e.g. based on clusters, and falling back on supervised FS strategies, aim to select features conducive to identify the pseudo labels (more in section 3). Eventually, unsupervised FS approaches are assessed within a supervised learning setting.

Following Y. LeCun's claim (LeCun, 2016) that unsupervised learning constitutes the bulk of machine learning, and that any feature can in principle define a learning goal, this paper tackles *Agnostic Feature Selection* with the goal of *leaving no feature behind*. Specifically, an unsupervised FS criterion aimed to select a subset of features supporting the prediction of *every* initial feature, is proposed. The proposed AGNOS approach combines AutoEncoders with structural regularizations, and delegates the combinatorial optimization problem at the core of feature selection to a regularized data compression scheme (section 2).

The contribution of the paper is threefold. Firstly, three regularization schemes are proposed and compared to handle the redundancy of the initial data representation. Informally, if the feature set includes duplicated features, the probability of selecting *one* copy of this feature should increase; but the probability of selecting several copies of any feature should be very low at all times. Several types of regularizations are proposed and compared to efficiently handle feature redundancy: regularization based on slack variables (AGNOS-S); L_2-L_1 regularization based on the AutoEncoder weights (AGNOS-W); and L_2-L_1 regularization based on the AutoEncoder gradients (AGNOS-G).

A second contribution is to show on the scikit-feature benchmark (Li et al., 2018) that AGNOS favorably compares with the state of the art (He et al., 2005; Zhao and Liu, 2007; Cai et al., 2010; Li et al., 2012) considering the standard assessment procedure. A third contribution is to experimentally show the brittleness of this standard assessment procedure, demonstrating that it does not allow one to reliably compare unsupervised FS approaches (section 5). The paper concludes with a discussion and some perspectives for further research.

Notations. In the following, the dataset is denoted $X \in \mathbb{R}^{n \times D}$, with n the number of samples and D the number of features. x_i (respectively f_j) denotes the i-th sample (resp. the j-th feature). The feature set is noted $F = (f_1, ..., f_D)$. $f_i(x_k)$ denotes the value taken by the i-th feature on the k-th sample.

2 AGNOS

The proposed approach relies on feature construction, specifically on AutoEncoders, to find a compressed description of the data. As said, feature construction does not comply with the requirement of interpretability. Therefore, AGNOS will use an enhanced learning criterion to retrieve the initial features most essential to approximate *all features*, in line with the goal of *leaving no feature behind*.

This section is organized as follows. For the sake of self-containedness, the basics of AutoEncoders are summarized in section 2.1. A key AutoEncoder hyper-parameter is the dimension of the latent representation (number of neurons in the hidden layer), which should be set according to the intrinsic dimension (ID) of the data for the sake of information preserving. Section 2.2 thus briefly introduces the state of the art in ID estimation.

In order to delegate the feature selection task to the AutoEncoder, the learning criterion is regularized to be robust w.r.t. redundant feature sets. A first option

considers weight-based regularization along the lines of LASSO (Tibshirani, 1996) and Group-LASSO (Yuan and Lin, 2007) (section 2.4). A second option uses a regularization defined on the gradients of the encoder ϕ (section 2.4). A third option uses slack variables, inspired from (Leray and Gallinari, 1999; Goudet et al., 2018) (section 2.5).

2.1 AutoEncoders

AutoEncoders (AE) are a class of neural networks designed to perform data compression via feature construction. The encoder ϕ and the decoder ψ are trained to approximate identity, i.e. such that for each training point x

$$\psi \circ \phi(x) \approx x$$

in the sense of the Euclidean distance, where the dimension d of the hidden layer is chosen to avoid the trivial solution of $\phi = \psi = Id$. Formally,

$$\phi, \psi = \arg\min \sum_{i=1}^{n} \|x_i - \psi \circ \phi(x_i)\|_2^2$$

Letting f_i denote the i-th initial feature and $\hat{f}_i$ its reconstructed version, the mean square error (MSE) loss above can be rewritten as :

$$L(F) = \sum_{i=1}^{D} ||\hat{f}_i - f_i||_2^2 \tag{1}$$

The use of AE to support feature selection raises two difficulties. The first one concerns the setting of the dimension d of the hidden layer (more below). The second one is the fact that the MSE loss (Eq. 1) is vulnerable to the redundancy of the initial description of the domain: typically when considering duplicated features, the effort devoted by the AE to the reconstruction of this feature increases with its number of duplicates. In other words, the dimensionality reduction criterion is biased to favor redundant features.

2.2 Intrinsic dimension

The *intrinsic dimension* (ID) of a dataset is informally defined as *the minimal number of features necessary to represent the data without losing information.* Therefore, a necessary (though not sufficient) condition for the auto-encoder to preserve the information in the data is that the hidden layer is at least as large as the ID of the dataset. Many different mathematical formalizations of the concept of ID were proposed over the years, e.g. Hausdorff dimension (Gneiting et al., 2012) or box counting dimension (Falconer, 2004). Both the ML and statistical physics communities thoroughly studied the problem of estimating the ID of a dataset empirically (Levina and Bickel, 2005; Camastra and Staiano, 2016; Facco

et al., 2017), notably in relation with data visualization (Maaten and Hinton, 2008; McInnes et al., 2018).

The best known linear ID estimation relies on Principal Component Analysis, considering the eigenvalues λ_i (with $\lambda_i > \lambda_{i+1}$) of the data covariance matrix and computing d such that the top-d eigenvalues retain a sufficient fraction τ of the data inertia ($\sum_{i=1}^{d} \lambda_i^2 = \tau \sum_{i=1}^{D} \lambda_i^2$). Another approach is based on the minimization of the stress (Cox and Cox, 2000), that is, the bias between the distance of any two points in the initial representation, and their distance along a linear projection in dimension d. Non-linear approaches such as Isomap (Tenenbaum et al., 2000) or Locally Linear Embedding (Saul and Roweis, 2003), respectively aim at finding a mapping on $\mathbb{R}^d$ such that it preserves the geodesic distance among points or the local barycentric description of the data.

The approach used in the following relies instead on the Poisson model of the number of points in the hyper-sphere in dimension d $\mathcal{B}(0, r)$, increasing like r^d (Levina and Bickel, 2005; Facco et al., 2017). Considering for each point x its nearest neighbor x' and its 2nd nearest neighbor x", defining the ratio $\mu(x) = \|x - x'\|/\|x - x"\|$ and averaging μ over all points in the dataset, it comes (Facco et al., 2017):

$$d = \frac{\log(1 - H(\mu))}{\log(\mu)} \tag{2}$$

with $log(1 - H(\mu))$ the linear function associating to $log(\mu_i)$ its normalized rank among the $\mu_1, \ldots \mu_n$ in ascending order.[1]

2.3 AgnoS

AgnoS proceeds like a standard AutoEncoder, with every feature being preliminarily normalized and centered. As the dimension of the latent representation is no less than the intrinsic dimension of the data by construction, and further assuming that the neural architecture of the AutoEncoder is complex enough, the AE defines a latent representation capturing the information of the data to the best possible extent (Eq. 1).

The key issue in AgnoS is twofold. The first question is to extract the initial features best explaining the latent features; if the latent features capture all the data information, the initial features best explaining the latent features will be sufficient to recover *all* features. The second question is to address the feature redundancy and prevent the AE to be biased in favor of the most redundant features.

Two approaches have been considered to address both goals. The former one is inspired from the well-known LASSO (Tibshirani, 1996) and Group-LASSO

[1] That is, assuming with no loss of generality that

$$\mu_i < \mu_{i+1}$$

one approximates the curve $(log(1 - i/n), log(\mu_i))$ with a linear function, the slope of which is taken as approximation of d.

(Yuan and Lin, 2007). These approaches are extended to the case of neural nets (below). The latter approach is based on a particular neural architecture, involving slack variables (section 2.5). In all three cases, the encoder weight vector W is normalized to 1 after each training epoch ($\|W\|_2 = 1$), to ensure that the LASSO and slack penalizations are effective.

2.4 AGNOS with LASSO regularization

This section first provides a summary of the basics of LASSO and group-LASSO. Their extension within the AutoEncoder framework to support feature selection is thereafter described.

LASSO and Group-LASSO Considering a standard linear regression setting on the dataset $\{(x_i, y_i), x_i \in \mathbb{R}^D, y_i \in \mathbb{R}, i = 1 \dots n\}$, the goal of finding the best weight vector $\beta \in \mathbb{R}^D$ minimizing $\sum_i \|y_i - \langle x_i, \beta \rangle\|^2$ is prone to overfitting in the large D, small n regime. To combat the overfitting issue, Tibshirani (1996) introduced the *LASSO* technique, which adds a L_1 penalization term to the optimization problem, parameter $\lambda > 0$ governing the severity of the penalization:

$$\beta^* = \arg\min_{\beta} ||\mathbf{y} - X\beta||_2^2 + \lambda||\beta||_1 \tag{3}$$

Compared to the mainstream L_2 penalization (which also combats overfitting), the L_1 penalization acts as a sparsity constraint: every i-th feature with corresponding weight $\beta_i = 0$ can be omitted, and the L_1 penalization effectively draws the weight of many features to 0. Note that in counterpart the solution is no longer rotationally invariant (Ng, 2004).

The *group LASSO* (Yuan and Lin, 2007) and its many variants (Meier et al., 2008; Simon et al., 2013; Ivanoff et al., 2016) have been proposed to retain the sparsity property while preserving the desired invariances among the features. Let us consider a partition of the features in groups $G_1, \dots, G_k$, the L_2-L_1 penalized regression setting reads:

$$\beta^* = \arg\min_{\beta} ||y - X\beta||_2^2 + \lambda \sum_{i=1}^{k} \frac{1}{|G_i|} \sqrt{\sum_{j \in G_i} \beta_j^2} \tag{4}$$

where the L_1 part enforces the sparsity at the group level (as many groups are inactive as possible) while preserving the rotational invariance within each group.

AGNOS-W: with L_2-L_1 weight regularization Under the assumption that all latent variables are needed to reconstruct the initial features (section 2.2), denoting $\phi(F) = (\phi_1 \dots, \phi_d)$ the encoder function, with $\phi_k = \sigma(\sum_{\ell=1}^{D} W_{\ell,k} f_\ell + W_{0,k})$ and $W_{i,j}$ the encoder weights, the impact of the i-th feature on the latent

variables is visible through the weight vector $W_{i,.}$. It thus naturally comes to define the L_2-L_1 penalization within the encoder as:

$$L(W) = \sum_{i=1}^{D} \sqrt{\sum_{k=1}^{d} W_{i,k}^2} = \sum_{i=1}^{D} \|W_{i,.}\|_2$$

and the learning criterion of AgnoS-W (Alg. 1) is accordingly defined as:

$$L(F) = \sum_{i=1}^{D} ||\hat{f}_i - f_i||_2^2 + \lambda L(W) \tag{5}$$

with λ the penalization weight. The sparsity pressure exerted through penalty $L(W)$ will result in setting $W_{i,.}$ to 0 whenever the contribution of the i-th initial variable is not necessary to reconstruct the initial variables, that is, when the i-th initial variable can be reconstructed from the other variables.

This learning criterion thus expectedly supports the selection of the most important initial variables. Formally, the score of the i-th feature is the maximum absolute value of $W_{i,j}$ for j varying in $1, \dots D$:

$$Score_W(f_i) = \|W_{i,.}\|_\infty \tag{6}$$

The rationale for considering the infinity norm of $W_{i,.}$ (as opposed to its L_1 or L_2 norm) is based on the local informativeness of the feature (see also *MCFS* (Cai et al., 2010), section 3): the i-th feature matters as long as it has a strong impact on at least one of the latent variables.

The above argument relies on the assumption that all latent variables are needed, which holds by construction.[2]

Algorithm 1 AgnoS-W

Input : Feature set $F = \{f_1, ..., f_D\}$
Parameter : λ
Output : Ranking of features in F
Normalize each feature to zero mean and unit variance.
Estimate intrinsic dimension $\widehat{ID}$ of F.
Initialize neural network with $d = \widehat{ID}$ neurons in the hidden layer.
Repeat
Backpropagate $L(F) = \sum_{i=1}^{D} \|\hat{f}_i - f_i\|_2^2 + \lambda \sum_{i=1}^{D} \|W_{i,.}\|_2$
until convergence
Rank features by decreasing scores with $Score_W(f_i) = \|W_{i,.}\|_\infty$.

[2] A question however is whether all latent variables are equally important. It might be that some latent variables are more important than others, and if an initial variable f_i matters a lot for an unimportant latent variable, the f_i relevance might be low. Addressing this concern is left for further work.

AGNOS-G: with L_2-L_1 gradient regularization In order to take into account the overall flow of information from the initial variables f_i through the auto-encoder, another option is to consider the gradient of the encoder $\phi = (\phi_1 \dots \phi_d)$. Varga et al. (2017); Alemu et al. (2018); Sadeghyan (2018) have recently highlighted the benefits of hidden layer gradient regularization for improving the robustness of the latent representation.

Along this line, another L_2-L_1 regularization term is considered:

$$L(\phi) = \sum_{i=1}^{D} \sqrt{\sum_{k=1}^{n} \sum_{j=1}^{d} \left(\frac{\partial \phi_j}{\partial f_i}(x_k) \right)^2}$$

and the learning criterion is likewise defined as:

$$L(F) = \sum_{i=1}^{D} ||\hat{f}_i - f_i||_2^2 + \lambda L(\phi) \tag{7}$$

The sparsity constraint now pushes toward cancelling all gradients of the ϕ_j w.r.t. an initial variable f_i. The sensitivity score derived from the trained auto-encoder, defined as:

$$\text{Score}_G(f_i) = \max_{1 \leq j \leq d} \sum_{k=1}^{n} \left(\frac{\partial \phi_j}{\partial f_i}(x_k) \right)^2 \tag{8}$$

is used to rank the features by decreasing score. The rationale for using the max instead of the average is same as for Score_W. Note that in the case of an encoder with a single hidden layer with *tanh* activation, one has:

$$\text{Score}_G(f_i) = \max_{1 \leq j \leq d} \sum_{k=1}^{n} \left(W_{i,j}(1 - \phi_j(x_k)^2) \right)^2 \tag{9}$$

Algorithm 2 AGNOS-G

Input : Feature set $F = \{f_1, ..., f_D\}$
Parameter : λ
Output : Ranking of features in F
Normalize each feature to zero mean and unit variance.
Estimate intrinsic dimension $\widehat{ID}$ of F.
Initialize neural network with $d = \widehat{ID}$ neurons in the hidden layer.
Repeat
Backpropagate $L(F) = \sum_{i=1}^{D} ||\hat{f}_i - f_i||_2^2 + \lambda \sum_{i=1}^{D} \sqrt{\sum_{k=1}^{n} \sum_{j=1}^{d} \left(\frac{\partial \phi_j}{\partial f_i}(x_k) \right)^2}$
until convergence
Rank features by decreasing scores with $Score_G(f_i) = \max_{j \in [1,\dots,d]} \sum_{k=1}^{n} \left(\frac{\partial \phi_j}{\partial f_i}(x_k) \right)^2$.

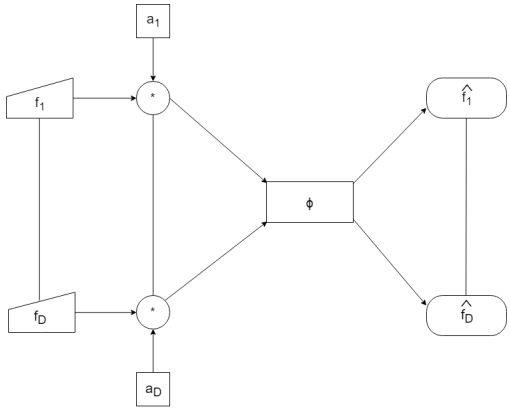

Fig. 1. Structure of the neural network used in AGNOS-S

2.5 AGNOS-S: with slack variables

A third version of AGNOS is considered, called AGNOS-S and inspired from Leray and Gallinari (1999); Goudet et al. (2018). The idea is to augment the neural architecture of the auto-encoder with a first layer made of *slack variables*. Formally, to each feature f_i is associated a (learned) coefficient a_i in [0,1], and the encoder is fed with the vector $(a_i f_i)$ (Fig. 1). The learning criterion here is the reconstruction loss augmented with an L_1 penalization on the slack variables:

$$L(F) = \sum_{i=1}^{D} ||\hat{f}_i - f_i||_2^2 + \lambda \sum_{i=1}^{D} |a_i| \tag{10}$$

Like in LASSO, the L_1 penalization pushes the slack variables toward a sparse vector. Eventually, the score of the i-th feature is set to $|a_i|$. This single valued coefficient reflects the contribution of f_i to the latent representation, and its importance to reconstruct the whole feature set.

Algorithm 3 AGNOS-S

Input : Feature set $F = \{f_1, ..., f_D\}$
Parameter : λ
Output : Ranking of features in F
Normalize each feature to zero mean and unit variance.
Estimate intrinsic dimension $\widehat{ID}$ of F.
Initialize neural network with $(a_1, ..., a_D) = \mathbf{1_D}$ and $d = \widehat{ID}$ neurons in the hidden layer.
Repeat
Backpropagate $L(F) = \sum_{i=1}^{D} ||\hat{f}_i - f_i||_2^2 + \lambda \sum_{i=1}^{D} |a_i|$
until convergence
Rank features by decreasing scores with $Score_S(f_i) = |a_i|$.

3 Related work

This section briefly presents related work in unsupervised feature selection. We then discuss the position of the proposed AGNOS.

Most unsupervised FS algorithms rely on *spectral clustering theory* (Luxburg, 2007). Let sim and M respectively denote a similarity metric on the instance space, e.g. $sim(x_i, x_j) = exp\{-\|x_i - x_j\|_2^2\}$ and M the $n \times n$ matrix with $M_{i,j} = sim(x_i, x_j)$. Let Δ be the diagonal degree matrix associated with M, i.e. $\Delta_{ii} = \sum_{k=1}^{n} M_{ik}$, and $L = \Delta^{-\frac{1}{2}}(\Delta - M)\Delta^{-\frac{1}{2}}$ the normalized Laplacian matrix associated with M.

Spectral clustering relies on the diagonalization of L, with λ_i (resp. ξ_i) the eigenvalues (resp. eigenvectors) of L, with $\lambda_i \leq \lambda_{i+1}$. Informally, the ξ_i are used to define soft cluster indicators (i.e. the degree to which x_k belongs to the i-the cluster being proportional to $\langle x_k, \xi_j \rangle$), with λ_k measuring the inter-cluster similarity (the smaller the better).

The general unsupervised clustering scheme proceeds by clustering the samples and falling back on supervised feature selection by considering the clusters as if they were classes; more precisely, the features are assessed depending on how well they separate clusters. Early unsupervised clustering approaches, such as the *Laplacian score* (He et al., 2005) and *SPEC* (Zhao and Liu, 2007), score each feature depending on its average alignment with the dominant eigenvectors ($\langle f_i, \xi_k \rangle$).

A finer-grained approach is *MCFS* (Cai et al., 2010), that pays attention to the local informativeness of features and evaluates features on a per-cluster basis. Each feature is scored by its maximum alignment over the set of eigenvectors ($max_k \langle f_i, \xi_k \rangle$).

Letting A denote the feature importance matrix, with $A_{i,k}$ the relevance score of f_i for the k-th cluster, *NDFS* (Li et al., 2012) aims to actually reduce the number of features. The cluster indicator matrix Ξ (initialized from eigenvectors $\xi_1, \ldots, \xi_n$) is optimized jointly with the feature importance matrix A, with a sparsity constraint on the rows of A (few features should be relevant).

SOGFS (Nie et al., 2016) goes one step further and also learns the similarity matrix. After each learning iteration on Ξ and A, M is recomputed where the distance/similarity among the samples is biased to consider only the most important features according to A.

Discussion. A first difference between the previous approaches and the proposed AGNOS, is that the spectral clustering approaches (with the except of Nie et al. (2016)) rely on the Euclidean distance between points in $\mathbb{R}^D$. Due to the curse of dimensionality however, the Euclidean distance in high dimensional spaces is notoriously poorly informative, with all samples being far from each other (Duda et al., 2012). Quite the contrary, AGNOS builds upon a non-linear dimensionality reduction approach, mapping the data onto a low-dimensional space.

Another difference regards the robustness of the approaches w.r.t. the redundancy of the initial representation of the data. Redundant features can indeed

distort the distance among points, and thus bias spectral clustering methods, with the except of Li et al. (2012); Nie et al. (2016). In practice, highly correlated features tend to get similar scores according to *Laplacian score*, *SPEC* and *MCFS*. Furthermore, the higher the redundancy, the higher their score, and the more likely they will *all* be selected. This weakness is addressed by *NDFS* and *SOGFS* via the sparsity constraint on the rows of A, making it more likely that only one out of a cluster of redundant features be selected.

Finally, a main difference between the cited approaches and ours is the ultimate goal of feature selection, and the assessment of the methods. As said, unsupervised feature selection methods are assessed along a supervised setting: considering a target feature f^* (not in the feature set), the FS performance is measured from the accuracy of a classifier trained from the selected features. This assessment procedure thus critically depends on the relation between f^* and the features in the feature set. Quite the contrary, the proposed approach aims to data compression; it does not ambition to predict some target feature, but rather to approximate *every* feature in the feature set.

4 Experimental setting

4.1 Goal of experiments

Our experimental objective is threefold: we aim to compare the three versions of AgnoS to unsupervised FS baselines w.r.t. i) supervised evaluation; and ii) data compression. Thirdly, these experiments will serve to confirm or infirm our claim that the typical supervised evaluation scheme is unreliable.

4.2 Experimental setup

Experiments are carried on eight datasets taken from the scikit-feature database (Li et al., 2018), an increasingly popular benchmark for feature selection (Chen et al., 2017). These datasets include face image, sound processing and medical data. In all datasets but one (Isolet), the number of samples is small w.r.t. the number of features D. Dataset size, dimensionality, number of classes, estimated intrinsic dimension[3] and data type are summarized in Table 1. The fact that the estimated ID is small compared to the original dimensionality for every dataset highlights the potential of feature selection for data compression.

AgnoS-W, AgnoS-G and AgnoS-S are compared to four unsupervised FS baselines : the *Laplacian score* (He et al., 2005), *SPEC* (Zhao and Liu, 2007), *MCFS* (Cai et al., 2010) and *NDFS* (Li et al., 2012). All implementations have also been taken from the scikit-feature database, and all their hyperparameters have been set to their default values. In all experiments, the three variants of AgnoS are ran using a single hidden layer, *tanh* activation for both encoder

[3] The estimator from Facco et al. (2017) was used as this estimator is empirically less computationally expensive, requires less datapoints to be accurate, and is more resilient to high-dimensional noise than other ID estimators (section 2.2).

Table 1. Summary of benchmark datasets

	# samples	# features	# classes	Estimated ID	Data type
arcene	200	10000	2	40	Medical
Isolet	1560	617	26	9	Sound processing
ORL	400	1024	40	6	Face image
pixraw10P	100	10000	10	4	Face image
ProstateGE	102	5966	2	23	Medical
TOX171	171	5748	4	15	Medical
warpPie10P	130	2400	10	3	Face image
Yale	165	1024	15	10	Face image

and decoder, *Adam* (Kingma and Ba, 2015) adjustment of the learning rate, initialized to 10^{-2}. Dimension d of the hidden layer is set for each dataset to its estimated intrinsic dimension $\widehat{ID}$. Conditionally to $d = \widehat{ID}$, preliminary experiments have shown a low sensitivity of results w.r.t. penalization weight λ in the range $[10^{-1}, ..., 10^{1}]$, and degraded performance for values of λ far outside this range in either direction. Therefore, the value of λ is set to 1. The AE weights are initialized after Glorot and Bengio (2010). Each performance indicator is averaged on 10 runs with same setting; the std deviation is negligible (Doquet, To appear in 2019).

For a given benchmark dataset, unsupervised FS is first performed with the four baseline methods and the three AGNOS variants, each algorithm producing a ranking S of the original features. Two performance indicators, one *supervised* and one *unsupervised*, are then computed to assess and compare the different rankings.

Following the typical supervised evaluation scheme, the first indicator is the K-means *clustering accuracy (ACC)* (He et al., 2005; Cai et al., 2010) for predicting the ground truth target f^*. In the following, clustering is performed considering the top $k = 100$ ranked features w.r.t. S, with $K = c$ clusters, with c the number of classes in f^*.

The second indicator corresponds to the unsupervised FS goal of recovering *every* initial feature f. For each $f \in F$, a 5-NearestNeighbor regressor is trained to fit f, where the neighbors of each point x are computed considering the Euclidean distance based on the top $k = 100$ ranked features w.r.t. S. The goodness-of-fit is measured via the R^2 *score* (a.k.a. *coefficient of determination*) $R^2(f, S) \in]-\infty, 1]$. The unsupervised performance of S is the individual R^2 score averaged over the *whole* feature set F (the higher the better):

$$Score(S) = \frac{1}{D} \sum_{j=1}^{D} R^2(f_j, S)$$

Table 2. Clustering ACC score on the ground truth labels, using the top 100 ranked features. Statistically significantly (according to a t-test with a p-value of 0.05) better results in boldface

	Arcene	Isolet	ORL	pixraw10P	ProstateGE	TOX171	warpPIE10P	Yale
AgnoS-S	**0.665**	0.536	**0.570**	**0.812**	**0.608**	0.404	0.271	0.509
AgnoS-W	0.615	**0.583**	0.548	0.640	0.588	0.292	0.358	0.382
AgnoS-G	0.630	0.410	0.528	0.776	0.569	0.357	**0.419**	**0.533**
Laplacian	0.660	0.482	0.550	0.801	0.578	0.450	0.295	0.442
MCFS	0.550	0.410	0.562	0.754	0.588	**0.480**	0.362	0.400
NDFS	0.510	0.562	0.538	0.783	0.569	0.456	0.286	0.442
SPEC	0.655	0.565	0.468	0.482	0.588	0.474	0.333	0.400

5 Experimental results and discussion

Supervised FS assessment. Table 2 reports the ACC score for each selection method and dataset. On all datasets but *TOX171*, the highest ACC is achieved by one of the three AgnoS variants, showing the robustness of AgnoS compared with the baselines. On average, AgnoS-S outperforms AgnoS-W and AgnoS-G.

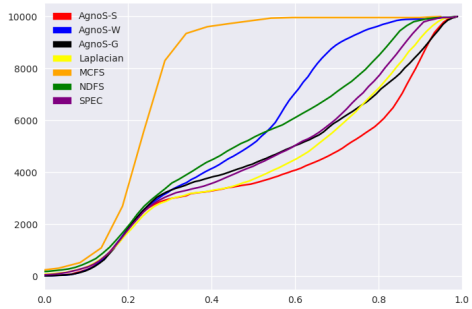

Fig. 2. Cumulative distribution functions of the R^2 scores of a 5-NearestNeighbors regressor using the top 100 ranked features on *Arcene*. If a point has coordinates (x, y), then the goodness-of-fit of the regressor is $\leq x$ for y initial features

Data compression FS assessment. Fig. 2 depicts the respective cumulative distribution of the R^2 scores for all selection methods on the *Arcene* dataset. A first observation is that every FS algorithm leads to accurate fitting (R^2 score > 0.8) for *some* features and poor fitting (R^2 score < 0.2) on *some* other features. This empirical evidence suggests that the prediction based on the selected features is very sensitive w.r.t. the variable to predict, supporting our claim that supervised assessment of unsupervised FS (dealing with a *single* target) is unreliable.

Another observation is that FS algorithms differ in the number of poorly fitted features. R^2 scores < 0.2 are achieved for less than 20% of features using any declination of AgnoS and more than 35% of features using MCFS, showing that AgnoS retains information about more features than MCFS on the *Arcene* dataset.

Table 3. Average of R^2 score of 5-NearestNeighbors regressor fitting *any* feature, using the top 100 ranked features. Statistically significantly (according to a t-test with a p-value of 0.05) better results in boldface

	Arcene	Isolet	ORL	pixraw10P	ProstateGE	TOX171	warpPIE10P	Yale
AgnoS-S	**0.610**	**0.763**	**0.800**	**0.855**	**0.662**	**0.581**	**0.910**	**0.703**
AgnoS-W	0.460	0.762	0.795	0.782	0.620	0.580	0.897	0.696
AgnoS-G	0.560	0.701	0.780	0.832	0.606	0.528	0.901	0.671
Laplacian	0.576	0.680	0.789	0.840	0.655	0.563	0.903	0.601
MCFS	0.275	0.720	0.763	0.785	0.634	0.549	0.870	0.652
NDFS	0.490	0.747	0.796	0.835	0.614	0.520	0.904	0.677
SPEC	0.548	0.733	0.769	0.761	0.646	0.559	0.895	0.659

Table 3 reports the average R^2 score of a 5-NearestNeighbors regressor on the whole feature set, for each FS algorithm and dataset. AgnoS-S is shown to achieve a higher mean R^2 score than AgnoS-W, AgnoS-G and all baselines on all datasets. These results empirically demonstrate that the selection subsets induced by AgnoS-S retain more information about the features *on average* than the baselines.

Notably, AgnoS-S generally outperforms AgnoS-W and AgnoS-G in a very significant manner, while AgnoS-W and AgnoS-G happen to be outperformed by the baselines. A tentative interpretation for this difference of performance among the three AgnoS variants is based on the key difference between the LASSO regularization and the slack variables. On one hand, the encoder weights in AgnoS-W (resp. the encoder gradients in AgnoS-G) are simultaneously responsible for producing the compressed data representation and enforcing sparsity among the original features. On the other hand, the slack variables in AgnoS-S are only subject to the sparsity pressure exerted by the L_1 penalty and have no other functional role. It is thus conjectured that the optimization of the slack variables can enforce sparse feature selection more efficiently than in AgnoS-W and AgnoS-G.

Sensitivity w.r.t. the number of selected features. Fig. 3 reports the R^2 score (averaged on the whole feature set) achieved by a 5-NearestNeighbors regressor on the *Yale* dataset for a number k of selected features in $[5, 10, \ldots, 200]$. AgnoS-S is shown to reliably outperform the baselines for every value of k (with the exception of $k \in \{5, 10\}$ where it is tied with NDFS).

Additionally, the *unsupervised* ranking of the considered FS algorithms appears to be stable w.r.t. k. This stability property does not hold using the ACC

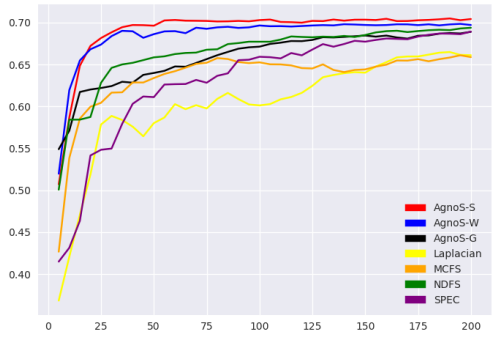

Fig. 3. Average R^2 score on *Yale* w.r.t. the number k of top ranked features considered

score, for which additional experiments have shown that the *supervised* ranking of FS algorithms is sensitive w.r.t. k (Doquet, To appear in 2019), confirming again the brittleness of the mainstream supervised assessment of feature selection methods.

Table 4. Empirical runtimes on a single Nvidia Geforce GTX 1060 GPU, in seconds

	arcene	Isolet	ORL	pixraw10P	ProstateGE	TOX171	warpPie10P	Yale
AgnoS	265	25	29	242	145	143	31	14
Laplacian	<1	<1	<1	<1	<1	<1	<1	<1
SPEC	3	9	<1	2	1	2	1	<1
MCFS	<1	2	<1	<1	<1	<1	<1	<1
NDFS	130	16	17	193	80	76	18	7

A main limitation of the proposed approach is its computational time. Table 4 reports the empirical runtimes of the baselines and AgnoS. AgnoS is shown to be between 25% and 100% slower than NDFS, and several orders of magnitude slower than Laplacian score, SPEC and NDFS.

6 Conclusion and Perspectives

In this paper, we have introduced a novel unsupervised FS algorithm based on data compression. A main merit of the proposed AgnoS-S is to better recover the whole feature set (and the target feature) compared to the baselines, in counterpart for its higher computational cost. A second contribution of the paper is to empirically show that the supervised assessment of unsupervised FS methods is hardly reliable.

This work opens two perspectives for further studies. The first one is concerned with early stopping of the AE, aimed to reduce the computational cost of AGNOS. Another direction is to consider Variational AutoEncoders (VAE) (Kingma and Welling, 2013) instead of plain AEs, likewise augmenting the VAE loss with an L_1 penalization to achieve feature selection; the expected advantage of VAEs would be to be more robust when considering small datasets.

Acknowledgments

We wish to thank Diviyan Kalainathan for many enjoyable discussions. We also thank the anonymous reviewers, whose comments helped to improve the experimental setting and the assessment of the method.

References

H. Alemu, W. Wu, and J. Zhao. Feedforward neural networks with a hidden layer regularization method. *Symmetry*, 10(10), 2018.

D. Cai, C. Zhang, and X. He. Unsupervised feature selection for multi-cluster data. *International conference on Knowledge Discovery and Data mining*, pages 333–342, 2010.

F. Camastra and A. Staiano. Intrinsic dimension estimation: Advances and open problems. *Information Sciences*, 328:26–41, 2016.

J. Chen, M. Stern, M. J. Wainwright, and M. I. Jordan. Kernel feature selection via conditional covariance minimization. *In Advances in Neural Information Processing Systems*, pages 6946–6955, 2017.

T. F. Cox and M. A. Cox. *Multidimensional scaling*. Chapman and hall/CRC, 2000.

G. Doquet. *Unsupervised Feature Selection*. PhD thesis, Université Paris-Sud, To appear in 2019.

F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv:1702.08608*, 2017.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. John Wiley and Sons, 2012.

E. Facco, M. d'Errico, A. Rodriguez, and A. Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Nature*, 7(1), 2017.

K. Falconer. *Fractal geometry: mathematical foundations and applications*. John Wiley and Sons, 2004.

X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. *International conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.

T. Gneiting, H. Ševčíková, and D. B. Percival. Estimators of fractal dimension: Assessing the roughness of time series and spatial data. *Statistical Science*, 27: 247–277, 2012.

I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *International Conference on Learning Representations*, 2015.

O. Goudet, D. Kalainathan, P. Caillou, I. Guyon, D. Lopez-Paz, and M. Sebag. Learning functional causal models with generative neural networks. *In Explainable and Interpretable Models in Computer Vision and Machine Learning*, pages 39–80, 2018.

X. He, D. Cai, and P. Niyogi. Laplacian score for feature selection. *Advances in Neural Information Processing Systems*, pages 507–514, 2005.

S. Ivanoff, F .Picard, and V. Rivoirard. Adaptive lasso and group-lasso for functional poisson regression. *The Journal of Machine Learning Research*, 17(1):1903–1948, 2016.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv:1312.6114*, 2013.

Y. LeCun. The next frontier in AI : Unsupervised learning. https://www.youtube.com/watch?v=IbjF5VjniVE, 2016.

P. Leray and P. Gallinari. Feature selection with neural networks. *Behaviormetrika*, 26(1):145–166, 1999.

E. Levina and P. J. Bickel. Maximum likelihood estimation of intrinsic dimension. *In Advances in Neural Information Processing Systems*, pages 777–784, 2005.

J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu. Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, 50(6), 2018.

Z. Li, Y. Yang, Y. Liu, X. Zhou, and H. Lu. Unsupervised feature selection using non-negative spectral analysis. *AAAI*, 2012.

U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

L. V. D. Maaten and G. Hinton. Visualizing data using t-SNE. *The Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv:1802.03426*, 2018.

L. Meier, S. Van De Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

A. Y. Ng. Feature selection, l_1 vs. l_2 regularization, and rotational invariance. *International Conference on Machine Learning*, 2004.

F. Nie, W. Zhu, and X. Li. Unsupervised feature selection with structured graph optimization. *AAAI*, pages 1302–1308, 2016.

S. Sadeghyan. A new robust feature selection method using variance-based sensitivity analysis. *arXiv:1804.05092*, 2018.

L. K. Saul and S. T. Roweis. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Journal of Machine Learning research*, 4(Jun): 119–155, 2003.

N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B (Methodological)*, pages 267–288, 1996.

D. Varga, A. Csiszárik, and Z. Zombori. Gradient regularization improves accuracy of discriminative models. *arXiv:1712.09936*, 2017.

T. Wiatowski and H. Bölcskei. A mathematical theory of deep convolutional neural networks for feature extraction. *IEEE Transactions on Information Theory*, 64(3):1845–1866, 2018.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2007.

Z. Zhao and H. Liu. Spectral feature selection for supervised and unsupervised learning. *International Conference on Machine Learning*, 2007.